图书情报工作
第59卷 第8期 2015年4月



# 起源技术在长期保存中的应用与研究

■ 吴振新[1] 李文燕[1,2]
[1]中国科学院文献情报中心 北京 100190 [2]中国科学院大学 北京 100049

**摘要：**[**目的/意义**]结合数据起源的内容和长期保存特点，全面研究和分析数据起源在长期保存中的应用，为长期保存系统组织管理起源提供参考。[**方法/过程**]分析长期保存领域中相关标准如OAIS、PREMIS和TRAC对起源的解释和要求，对比起源在已有的长期保存系统中的应用情况。[**结果/结论**]提出以事件为核心的长期保存起源管理框架，总结起源的详细内容、捕获方法、组织方案、存储封装策略和技术方案等。

**关键词：**起源 事件 长期保存 保存周期 实践
**分类号：**G250.76
**DOI：**10.13266/j.issn.0252-3116.2015.08.017

## 1 前言

起源(provenance,又译为溯源或来源)，代表了数字对象的产生及发展历史。通过记录起源信息，人们可以了解数字对象所发生的变化，以及变化的地点、原因、时间和责任人等7W[1](what、where、who、when、which、why、how)信息。起源记录的内容对于解决数据的可信性、结果可靠性、数据修改或分析过程的透明性以及数据引用来源等重要数据问题至关重要。

在长期保存系统中组织和管理起源具有重要意义。长期保存系统的首要目标是保证起源的真实性、可理解性和可访问性，如果数据失去了真实性，可理解性和可访问性也就无从谈起。真实性有两层含义：其一是原始内容没有发生改变；其二是原始内容发生了合理的变化。起源记录了长期保存系统对原始对象的各种操作，可以提供真实性判断的证据，显示数字发生了什么变化，对数字对象产生了什么影响，以此证明数字对象是否真实。

起源研究从类型上大致分为数据级起源(如数据库起源)和过程级起源(如工作流起源)[2]。国外对这两类起源有较多的研究和实践[3-4]。近些年，国内也开始关注起源，包括数据级起源数据理论和实践[2]、起源技术综述[5]和表达模型分析[6]等。其中以模型的研究居多，如针对OPM(Open Provenance Model)的安全性改进[7]、针对W3C PROV[8]标准的介绍和Web应用，但鲜有把过程级起源和特定领域相结合的研究和实践。

长期保存领域的起源属于过程级起源，国外已有较多保存系统对起源做了研究和实践，国内尚处于起步阶段[7]，研究和应用较少。在此背景下，本文首先分析相关标准对起源的要求，然后对比起源在长期保存系统中的应用，最后在此基础上提出以事件为核心的长期保存起源管理框架，并全面地分析该框架涉及的关键问题，包括功能流程、起源的内容、组织方案、存储策略和技术方案等内容，为起源技术在长期保存系统中的理论研究和实践应用提供借鉴和参考。

## 2 相关标准对起源的要求和描述

### 2.1 OAIS的要求与描述

OAIS(Open Archival Information System)是长期保存的基础框架，为长期保存提供了标准、规范化的保存系统功能流程和信息对象模型。OAIS对起源的定义、内容和作用等做了简要陈述。在OAIS中，起源被定义为内容信息的历史，展示了内容信息产生的由来、自产生以来所发生的变化和管理过程中保管责任方的变动[10]。以数字图书馆集合为例，其起源包含以下内容：①非原始数字内容：数字过程和主要版本链接。②数字出版物：原始版本链接、保存过程的元数据、更早版本的链接、改变历史和信息对象描述。

不同类型的数字对象的起源记录的内容类型并不相同，如对于空间科学数据，起源包括仪器信息、主要

作者简介：吴振新(ORCID:0000-0003-4966-1961)，研究馆员，硕士生导师；李文燕(ORCID:0000-0002-7695-5087)，硕士研究生，通讯作者，E-mail:liwenyan@mail.las.ac.cn。
收稿日期：2015-03-03 修回日期：2015-04-08 本文起止页码：118-125 本文责任编辑：王传清

研究者、软件接口规范信息等;对于软件包,起源包括修改历史、注册信息和版权等内容。

在长期保存系统中,原始数据以信息包为基本单位进行管理,包括 SIP(submission information package)、AIP(archival information package)和 DIP(dissemination information package)3 种类型。每个信息包由内容信息(content object)和保存描述信息(preservation description information,PDI)组成。内容信息是保存的原始目标,包含数字对象(data object)和表征信息(repretation information),PDI 负责解释内容信息。起源是 PDI 的重要组成部分,记录了数字对象的变化,为真实性判断、可信赖认证、信息审计、权限判断、版本变迁等提供重要依据。

起源也可以被看作一种元数据,但和其他元数据(如题名)不同的是起源信息是动态产生的。起源的产生贯穿了数字对象的整个生命周期:内容信息被摄入保存系统前,起源由内容生产者提供给长期保存系统;内容信息被摄入保存系统后,起源被保存系统不断地捕获,并更新到相关的模块。

## 2.2 PREMIS 的要求与描述

PREMIS(Preservation Metadata Implementation Strategies)是支持数字保存处理过程的信息框架,包括 PREMIS 数据字典[11]和 PREMIS 框架[12]两大部分。相对 OAIS,PREMIS 对起源做了更加详细的陈述,并定义了可用来描述起源的保存元数据语义单元。

PREMIS 数据字典指出,起源主要描述了责任人对数字对象保管和管理的责任、发生在数字对象生命周期内的关键事件,以及其他与数字对象的创建、管理和保存有关的信息。记录起源是保证数字对象可信赖的重要手段,可以从技术层面为真实性管理提供支持。所以在长期保存系统的配置过程中应特别注意起源的管理,并从数字对象完整生命周期角度有效地组织和维护起源。

PPREMIS 框架对起源作了更丰富和深刻的阐释。认为起源主要解释了内容数据对象从被创建开始到其当前状态过程中随时间迁移而发生的变化。除了记录内容对象的"时间表"之外,起源还是基于事件的元数据。换句话说,数字对象相关状态的演化是被重要事件驱动的,例如对象的创建、所有权的转移、被摄入存档系统的过程或对象的格式迁移都是由事件引起的。如图 1 所示,起源可以分为来源(origin)、摄入前期(pre-ingest)、摄入过程(ingest)、存档过程(archival retention)和权限管理(rights management)5 种类型,每种类型的内容都记录为事件。所以记录起源就可以转换为记录这些特定事件的细节,以及它们对内容数据对象的影响。

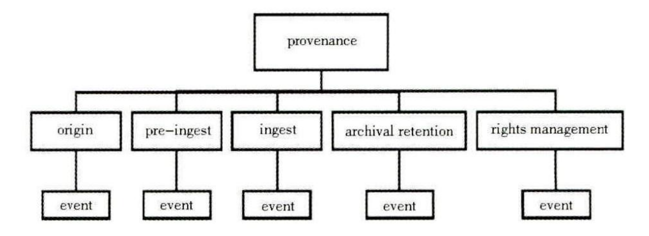



图 1 PREMIS 框架定义的起源事件模型[12]

PREMIS 数据字典定义了 5 种基本实体——语义实体(intellctual entity)、对象(object)、事件(event)、代理(agent)和权限声明(rights statement)。根据 PREMIS 框架对起源的解释,可以使用 Event 包含的语义单元来记录起源。不仅如此,从 OAIS 和 PREMIS 给出的定义来看,起源还应该包括操作过程中涉及的对象和责任者,如表 1 所示:

表 1 描述起源的 PREMIS 语义单元

| Event |
| --- |
| 2. 1 eventIdentifier |
| 2. 1. 1 eventIdentifierType |
| 2. 1. 2 eventIdentifierValue |
| 2. 2 eventType |
| 2. 3 eventDateTime |
| 2. 4 eventDetail |
| 2. 5 eventOutcomeInformation |
| 2. 5. 1 eventOutcome |
| 2. 5. 2 eventOutcomeDetail |
| 2. 5. 2. 1 eventOutcomeDetailNote |
| 2. 5. 2. 2 eventOutcomeDetailExtension |
| 2. 6 linkingAgentIdentifier |
| 2. 6. 1 linkingAgentIdentifierType |
| 2. 6. 2 linkingAgentIdentifierValue |
| 2. 6. 3 linkingAgentRole |
| 2. 7 linkingObjectIdentifier |
| 2. 7. 1 linkingObjectIdentifierType |
| 2. 7. 2 linkingObjectIdentifierValue |
| 2. 7. 3 linkingObjectRole |
| Object |
| 1. 10relationship |
| 1. 10. 3 relatedObjectIdentification |
| 1. 10. 3. 1 relatedObjectIdentifierType |
| 1. 10. 3. 2 relatedObjectIdentifierValue |
| 1. 10. 3. 3 relatedObjectSequence |
| 1. 10. 4 relatedEventIdentification |
| 1. 10. 4. 1 relatedEventIdentifierType |
| 1. 10. 4. 2 relatedEventIdentifierValue |
| 1. 10. 4. 3 relatedEventSequence |
| 1. 11 linkingEventIdentifier |
| 1. 11. 1 linkingEventIdentifierType |
| 1. 11. 2 linkingEventIdentifierValueAgent |
| 3. 1 agentIdentifier |
| 3. 1. 1 agentIdentifierType |
| 3. 1. 2 agentIdentifierValue |
| 3. 2 agentName |
| 3. 3 agentType |

其中，eventType 是受控词，是起源事件的重要内容，确定哪些类型的事件需要记录。哪些不要记录，是起源管理的重点，每个仓储都应该定义自己的 eventType 值。PREMIS 提供了一个事件类型清单供参考：①creation（新对象创建）；②deacession（从仓储目录中移除对象的过程）；③decompressionn（逆压缩）；④decryption（加密数据转换为明文）；⑤deletion（从仓储存储中移除一个对象）；⑥digital signature validation（确定解密的数字签名是否匹配期待值）；⑦dissemination（从仓储存储中检索一个对象，以为用户访问）；⑧fixity check（验证对象的在给定时期没有发生变化）；⑨ingestion（增加对象到保存仓储的过程）；⑩message digest calculation（摄入时生成原始数据的校验和过程）；⑪migration（对象创建新版本的转换）；⑫normalization（创建更有利于保存新版对象的转换）；⑬replication（创建对象副本的过程，与原数字对象比特流完全一致）；⑭validation（用标准对比一个对象的过程，没有任何不符合规范）；⑮virus check（检查文件是否收到恶意程序攻击）。

### 2.3 可信赖仓储认证标准的要求与描述

可信赖仓储认证标准（Trustworthy Repositories Audit & Certification，TRAC）即 ISO1636，为数字仓储库的真实性的审核和认证提供基础框架，是数字仓储库真实性审查的标准规范。

TRAC 中，对起源的描述主要出现在第 4 章和第 5 章，解释了起源发挥的作用、必要性和维护要求等。

起源提供复制和移动数据的过程信息，且必须被不断维护和升级，可以帮助确定责任人、数字对象副本的数量和位置。当 SIP 与 AIP（Archive Information Package）不一致时，仓储须根据书面规程进行处理，并且需要指明不一致的原因，起源可以发挥重要的作用。PDI 通过提供起源以及与其他信息之间的关联，确保内容信息能够被理解，这也是理解内容信息的关键元素。

根据协议，在数据对象处理过程中，除非协议另有说明，仓储可通过文档格式来判断保存对象的相关属性。在这种情况下，仓储需要对格式相同的保存对象的起源进行统一描述。为了使仓储拥有一套能够支持长期保存的 AIP 定义，必须能够识别和解析 AIP 中的必要组件。

因此，保存仓储需要有文档清晰地展示诸如表征信息和起源之类的 AIP 组件，使之能够被管理和及时更新。同时还要保存起源和 AIP 的关键信息如内容信息、表征信息和其他 PDI 的关联，并对它们之间的关联进行一致的定义。此外，为有效识别和解析起源，仓储还需要拥有一套机制来正确验证所有内容生产方的身份信息（起源信息的一部分），支持长期保存的 AIP 定义，并根据实际情况随着时间扩展起源。

## 3 起源在长期保存中的应用现状

### 3.1 DAITSS

DAITSS[13]是由佛罗里达图书馆自动化中心为佛罗里达数字保存系统（Florida Digital Archive，FDA）开发的一个数字保存仓储系统。它利用 METS 格式，把起源记录在管理元数据 amdSec 的 digiprovMD 元素里。对于起源，DAITSS 主要以事件的方式来记录，并把事件分为包级和文件级两个级别。包级事件包括 submit、ingest、disseminate、refresh 和 withdraw；文件级事件包括 virus check、describe、xml resolution、normalize 和 migrate。一个 AIP 的 METS 文件封装了 3 个级别的管理元数据，分别包含不同层次的起源：①第一级起源记录了协议信息；②第二级起源记录了 PREMIS 提交、摄入、分发、更新和撤销事件；③第三级起源记录了 DAITSS 服务为每个文件执行的 PREMIS 事件，如文件转换、病毒检查等。

### 3.2 CASPAR

CASPAR（Cultural，Artistic and Scientific Knowledge for Preservation，Access and Retrieval）[14]使用 IBM 开发的 PDS（Preservation DataStore）[15]来管理起源，主要用于权限管理、知识库更新跟踪和真实性管理。

在 PDS 中，起源被当作独立的信息对象处理，拥有自己的表征信息。一条起源记录就是一个起源事件。起源事件可分为内部事件和外部事件，内部事件可以被 PDS 自动捕获，外部事件则需要通过接口人工添加。起源数据的概念结构见图 2，起源数据由多个起源记录组成，每条起源记录包含唯一标示符（record ID），PDS 内部事件标志（PDS internal），事件内容（content）和表征信息（repInfo）4 部分。一个起源事件可能指向单个 AIP（如创建），或者一组 AIP（如包含某种数据的所有 AIP 被转换成某种新的格式），或者整个系统（如所有存档的所有者发生改变）。

### 3.3 APARSEN

APARSEN（Alliance Permanent Access to the Records of Science in Europe Network）[16]把起源作为真实性管理的主要证据来收集，并为此撰写一份详细的起源和真实性配置指导文档[17]，以实证其可行性。数字

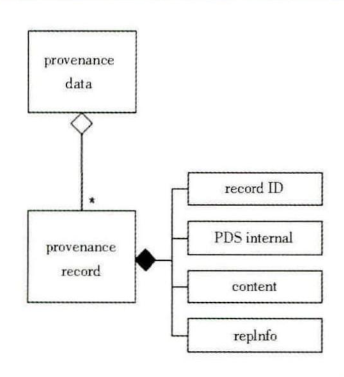



图2 CASPAR起源信息逻辑机构[14]

资源的生命周期被分为摄入前和保存期两个阶段,并通过事件记录起源。由于长期保存系统的复杂性,APARSEN把核心事件进行整理分类。每个阶段包括如下事件类型:①摄入前——捕获,整合,聚合,删除,迁移,转换,提交。②保存期——捕获,保存—摄入,保存—聚合,保存—抽取,保存—迁移,保存—删除,保存—转移。

这些事件可用以下元素进行描述:事件的描述、代理、输入、输出以及可信性证据记录(authenticity evidence record ,AER)。AER是能够作为真实性判断依据的起源信息。

APARSEN使用CRMdig[18]模型描述起源,并对$CRM_{dig}$和OPM做出了映射,增加其交互性。$CRM_{dig}$是一个以事件为核心来描述起源的本体模型,重点突出了对物理对象的起源的描述,对科学研究产生数字化数据的物理环境有丰富的描述。

### 3.4 SCAPE

SCAPE(Scalable Preservation Environments)[19]把起源应用到了SCAPE保存规划、数据出版平台和知识库模块。SCAPE使用Taverna定义和处理保存任务,并利用Taverna自带的起源插件Workbench 2.4输出Taverna工作流中的起源。Taverna拥有自己的起源本体—— tavernaprov,该本体扩展了本体PROV-O和wfprov,目的是描述一般模型无法表示的Taverna行为,如错误文档和迭代。与DATISS一样,SCAPE信息包也使用了METS文件的digiprovMD元素封装起源[20],起源由PREMIS事件和代理组成,并利用premis:object,premis:event和premis:agent相关语义单元描述起源。

### 3.5 其他相关项目

此外,长期保存的其他许多项目和系统也都涉及起源的收集和管理。iRODS(integrated Rule-Oriented Data System)[21]设计了分布式的起源信息系统,提供多结点的起源记录("P-Services")和起源查询(Q-Services)服务。记录的起源不仅包括内容数据和文件的变化历史,而且包括用户对文件访问、处理数据的规则版本变化和iRODS的系统信息等。PrestoPrime[22]通过事件和生产者或生产者代理来记录起源,事件被划分为存缴前事件和存缴事件,后者包括新版本产生和有效性检查,参考PREMIS字典,使用DNX和OPM两个模型来记录起源。Data Conservancy[23]把起源划分为起源服务和世系服务两部分,前者记录了发生在系统内的事件,后者记录了数据对象之间的关系,并通过HTTP调用Linage API和Event API两个Web接口来调用上述服务。

### 3.6 起源应用现状总结

起源的管理关键在于设计起源组织模型和记录流程。在对起源的组织方面,虽然不同保存系统使用了不同组织模型:如OPM、PREMIS、$CRM_{dig}$或自定义模型,但是却不约而同地以事件为核心来记录起源,这一点和PREMIS是一致的。除事件之外,还记录了生产者和对象关系这些重要信息。在对起源进行管理时,把起源作为元数据和内容信息一起保存,并按照一定的封装格式(如METS)进行组织。在起源组织管理方面,与技术元数据以及描述元数据不同,起源需要不断地更新,一条起源一般需要经过捕获、组织、封装存储等加工过程,并最终提供访问查询,或者被保存系统的其他模块调用。

## 4 基于OAIS框架的起源研究与分析

起源是PDI的重要组成部分,也是长期保存实践需要记录的重要内容,但是到目前为止,还缺乏对其综合、全面的分析。虽然OAIS和PREMIS等给出了概念定义和解释,但是具体记录什么内容、相关技术、起源管理策略等这些长期保存社区关心的问题并无明确的说明,其他项目也是针对自己项目的特点加以记录,还没有一个综合性的完整框架能为长期保存提供参考。本文将基于整个保存周期的角度对起源的内容、捕获、存储和封装进行全面的分析。

### 4.1 以事件为核心的起源信息管理框架与流程

基于对相关标准对起源的解释和以上项目分析,本文归纳并设计以事件为核心的起源信息管理框架。该框架是一个基于OAIS具有普适性的长期保存起源管理框架。如图3所示,中心部分是起源的管理模块,它嵌入在OAIS的功能模块中,动态地监测摄入、归档存储、数据管理、保存规划和业务管理各个流程的事

件。起源管理模块根据预先配置好的事件类型，在整个保存生命周期内捕获起源保存对象的起源事件，并按照相应事件起源模型组织成规范的起源信息，被单独或者和其他元数据保存在一起，以达到应用的目的。这个起源管理的过程是循环进行的，所以随着时间的推移，会不断地产生新的起源信息。

如图3所示，以事件为核心的起源的管理框架共涉及了4个基本流程模块，即捕获、组织、存储封装和访问，其中捕获、组织以及存储封装是重点。

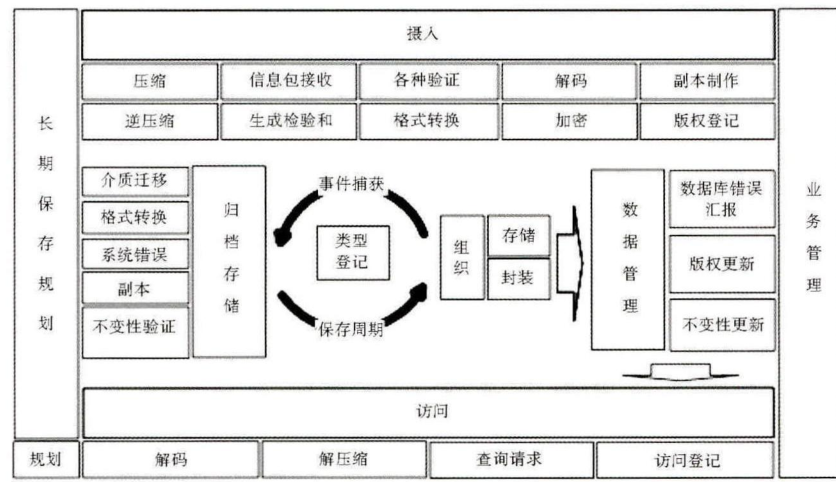



图3 长期保存的起源管理框架

捕获指的是根据事先设定好的事件类型，监控保存系统的相关内容，一旦事件发生，就将事件和必要信息记录下来，并通知组织模块。组织模块则按照设计好的的起源模型和元素，把捕获传来的内容记录为规范的起源信息。存储封装则负责将起源信息存储到相关的信息包文件或者数据库中，并对起源文件按照一定的格式封装保持其与数字对象之间的关联。访问模块主要提供用户查看和下载起源信息功能，或提供给长期保存系统中的其他保存模块（如真实性管理）使用。

## 4.2 起源的内容分析

虽然OAIS和PREMIS对起源作了解释，但是对具体记录什么，没有详细说明。在长期保存系统中，以事件为核心的起源主要记录了以下相关内容：

4.2.1 发生在数字对象上的事件 即对数字对象的操作，是起源的核心。事件内容包括事件标识符、事件发生时间、事件类型、事件细节、事件结果和事件使用的设备等。

4.2.2 关联的数字对象 即事件操作对象，包括输入对象和输出对象。应该完整保存事件涉及到的数字对象，通过在事件中引用数字对象的标识符将二者关联起来，但不包含数字对象的描述元数据。

4.2.3 相关代理 狭义的代理主要指操作人，长期保存系统有责任记录相关责任人在事件中发挥的作用，以判断这些操作是否合法。更广义的代理还包括组织、个人或者软件产品等。

4.2.4 对象之间的关系 即不同版本对象之间的关联。对象在被提交到保存系统后，会有多个副本或者不同格式的版本。在向用户发送DIP时，需要对其加以说明，原始版本是否发生变化，如果变化，提供现存版本，并加以说明。

其中，起源事件是起源核心，贯穿了从摄入开始至对象消亡的整个长期保存过程。笔者参考PREMIS，从对象的创建、修改和删除以及验证4个角度，列出了如图3所示事件：信息包接收、压缩、逆压缩、生成校验和、格式转换、解码、加密、副本制作、版权登记、介质迁移、系统错误、不变形验证、数据库错误、版权更新、保存规划、查询和访问登记。

## 4.3 起源模型的应用

描述起源的模型较多，典型的有PROV-DM[24]、OPM和Provenir。一些元数据方案、本体词汇等也提供了表达Provence的术语，如DC元数据、VoID词汇和Provence Vocabulary，国内均有相关研究[6]。虽然不同模型对起源的元素定义差异较大，但对起源的基本元素的定义却基本是一致的，即可识别的对象、处理对象的过程和涉及到的责任方。

在长期保存领域中，起源的组织方式经历了这样的变化。较早的做法，如英国国家归档中心（The National Archives，TNA）[25]自定义元数据，把保管历史、保管人、保存历史事件等字段映射为Provenance，起源分散在多个不同的表结构中，不利于整合。后来较为通用的做法是直接使用PREMIS事件实体作为起源，例如DAITSS和大英图书馆的数字图书馆产期保存项目。同期，CASPAR项目则使用$CRM_{dig}$来组织起源。为了增强起源的交互性和表述能力，PrestoPrime开始使用通用模型OPM和PREMIS两种方式采集不同的起源。近期，SCAPE则依赖Taverna的起源插件[26]，使用扩展的W3C标准PROV模型记录起源。由此也看出，长期保存社区对起源信息的日益重视，在组织起源方面，语义和结构越来越丰富，交互性和通用性越来越强。

### 4.4 起源的捕获方法和工具

起源的捕获方法主要有注释和逆置[5]。注释法实施比较简单,只需要记录下与数据相关的处理信息即可。工作流起源管理系统一般都是采取注释方式。逆置法认为在一定的限制条件下,可以通过分析数据库操作语句得出任意粒度的起源。在长期保存系统中,常采用第一种方法来记录起源,通过多种途径实现:

(1)在系统内部编写独立函数模块、接口或组件。如 CASPAR 项目的 PDS,它由 IBM 开发,能够感知所有发生在 PDS 内的事件并记录为起源,同时提供方法使用户可以记录 PDS 无法感应的外部事件。

(2)在系统内嵌入起源捕获插件。如 SCAPE 使用了 Taverna 工作流软件,执行转换、迁移、副本制作等任务,抽取出工作流中的起源。此类插件一般使用通用起源模型组织起源,例如 W3C 的 PROV 和 OPM 等。此外,其他大多数的工作流系统,如 VisTrails、REDUX 和 VDS 都集成了起源捕获的功能,通过集成这些工作流软件,利用其提供的组块或插件有效捕捉系统内部的起源信息。

(3)使用元数据工具抽取起源信息,如 JHOVE、DROID 等。这类软件可以记录一部分起源,不能追踪完整的起源,这是目前捕获起源的一种常见方式。另外,也有专门针对起源的元数据抽取工具,例如由 Ex-Libris[27]开发的商业软件,它可以追踪到数字对象的改变历史,并保存为 PREMIS 事件。

(4)集成已有的起源引擎到保存系统来管理起源。如 iRODS 在系统中集成了开源起源管理引擎 PA-SOA,通过基于 Web 服务的接口实现对起源的管理。

遗憾的是,目前还没有专门针对长期保存系统的起源捕捉插件或工具包。虽然长期保存系统各不相同,但是大多遵守 OAIS 参考模型流程模块和信息对象模型,在此共同的基础上,或可以开发具有以下特点的开源工具包或插件:①能够使用 XML 或者数据库等方式灵活地配置和更新起源事件类型列表。②当起源事件库中的事件被触发时,相应模块能捕捉到事件并记录下来。③从捕获的事件中抽取出起源保存到数据库表或文件中。

### 4.5 起源的存储

存储起源有两种策略:一种方式是将起源和内容对象以及其他 PDI 信息一起保存。在 METS 文件中,起源信息和版权信息被保存在管理元数据区域,这种混合存储的方法,优点是易于维护起源的完整性,缺点是难于发布和检索。另一种方式是将所有的起源单独存储到一个数据库或者文件中。DataONE[28]把工作流起源存储在 Mysql 和图数据库构建的起源仓储中,目前还未和 DataONE 的保存仓储融合在一起。这种方式便于快速查询和可视化呈现起源,缺点是维护困难,当数据被修改时需要考虑起源版本变更等问题。

从存储的格式方面来看,一种是采用文件方式,如 XML、RDF 和 OWL 语言等。XML 是信息交换的主要格式,易兼容现有长期保存系统的元数据保存规范格式。随着关联数据和本体的发展,一些项目尝试使用语义化方式如 RDF,并在此基础上做出推理和查询。目前封装起源文件的格式有很多,被调研的项目中常用到的通用封装格式有 METS[30] 和 XFDU[31] 两种。METS 应用最为广泛,DAITSS、UK 期刊保存和 SCAPE 等项目都使用了这种封装格式,把起源封装在管理元数据的 <digprovMD> </ digprovMD >标签里。CASPAR 使用 XFDU 封装起源[32],它通过两种方式把起源封装在 metadataSection 位置,第一种在 XFDU 的 XML 文件中直接写入起源;第二种通过 URL 链接指向外部的起源文件。

除了文件方式之外,有的项目还使用数据库表来存储起源,如关系数据库 MySQL 和 Neo4j,这种方式查询效率较高,便于快速获取起源。

随着起源的累积,其容量可能会超过数据本身。实践过程中,可以把两种存储策略、多种格式相结合,把一部分起源,如不常用或更新频率低的起源,放在封装文件或者通过链接指向单独的文件,对于经常使用或更新频率较高的起源,直接存储在数据库表中,支持快速访问和查询。

## 5 总结与展望

以上对长期保存领域的数字对象的起源做了比较全面而深入的分析,但由于实践过程中的复杂环境和多样化的应用需求,还有一些问题需要深入考虑。

起源包括的范围十分广泛,并且与 PDI 的其他部分,如情境信息有所交叉,所以在在设计长期保存系统中起源的记录时,应该给予明确的界定,以免发生混淆。

在长期的保存过程中,数字对象会因各种保存管理而产生多种起源事件,需要对其进行全面的研究、分析和定义,并根据项目的实际目标和需求,制定一个全面的、个性化的起源管理规划,既记录足够的信息,充分保持被保存数据的完整性、真实性和可理解性,同时又不致产生过多的起源数据,增加保存系统维护的负

担。

与起源的组织、捕获和存储技术相比,对起源可视化技术的研究相对比较缺乏,有些系统至今只提供以XML文件的方式来访问起源,这种不友好的呈现方式既不利于阅读,也不能充分展现起源应有的价值,所以如何有效、多角度、生动化、清晰化地呈现起源,是长期保存中值得关注的事情。

同时,与内容数据一样,起源数据应该被妥善存储和保护。保存机构应该采取措施保护起源的安全性和真实性,在突发事件如数据变换、存档和转换过程中使用防篡改技术(如数字签名)以保护起源信息链的完整性、可靠性和有效性。

总地来说,起源对于数字对象真实性判断、版权归属、访问权限管理、知识库变迁等内容具有重要作用,它既是OAIS信息模型的一部分,又是长期保存系统实践中非常重要的支撑内容,所以长期保存系统应该充分地结合OAIS、PREMIS和TRAC等标准,根据自身的实际情况,制定出一套完善的起源应用管理方案。

下一步,将尝试在中国科学院文献情报中心的长期包系统(DPS)中集成起源的管理功能,并通过事件捕获起源:①根据系统的流程核定需要捕获的事件;②确定以PREMIS和PROV结合的方式记录起源的元数据方案;③设计起源的存储封装策略,包括存储方式、封装格式、存储位置等;④根据系统的技术平台选取合适的技术方案加以实施。最后从实践中验证和改进本文提出的以事件为核心的起源信息管理框架。

**参考文献:**

[1] Ram S, Liu J. A new perspective on semantics of data provenance [EB/OL]. [2015-03-01]. http://citeseerx. ist. psu. edu/viewdoc/download? doi=10. 1. 1. 154. 8485&rep=rep1&type=pdf.

[2] 王黎维,鲍芝峰,Koehler H,等. 一种优化关系型溯源信息存储的新方法[J]. 计算机学报,2011(10):1863-1875.

[3] Plale B, Gannon D, Simmhan Y L. A survey of data provenance techniques[EB/OL]. [2015-03-01]. http://citeseerx. ist. psu. edu/viewdoc/summary? doi=10. 1. 1. 70. 6294.

[4] Simmhan Y L, Plale B, Gannon D. A survey of data provenance techniques[J]. Computer Science Department, 2005,34(3):31-36.

[5] 戴超凡,王涛,张鹏程. 数据起源技术发展研究综述[J]. 计算机应用研究,2010(9):3215-3221.

[6] 沈志宏,张晓林. 语义网环境下数据溯源表达模型研究综述[J]. 现代图书情报技术,2011(4):1-8.

[7] 刘通. 基于OPM的安全起源研究[D]. 淄博:山东理工大学,2013.

[8] 倪静,孟宪学. PROV数据溯源模型及Web应用[J]. 图书情报工作,2014,58(3):13-19.

[9] 祝犇. 数字信息长期保存中来源感知技术的研究[D]. 武汉:华中科技大学,2013.

[10] CCSDS 650.0-M-2, Reference model for an open archival information system(OAIS)[S]. Washington: CCSDS,2012.

[11] PREMIS data dictionary for preservation metadata, version 2.0 [EB/OL]. [2015-03-01]. http://www. loc. gov/standards/premis/v2/premis-2-0. pdf.

[12] Preservation metadata and the OAIS information model: A metadata framework to support the preservation of digital objects, a report [M]. Dublin: OCLC/RLG Working Group on Preservation Metadata, 2002.

[13] The Florida Center for Library Automation. DAITSS website[EB/OL]. [2015-03-01]. http://daitss. fcla. edu/.

[14] Factor M, Henis E, Naor D, et al. Authenticity and provenance in long term digital preservation: Modeling and implementation in preservation aware storage[EB/OL]. [2015-03-01]. http://static. usenix. org/event/tapp09/tech/full_papers/factor/factor. pdf.

[15] IBM. Preservation dataStore interface[EB/OL]. [2015-03-01]. http://www. casparpreserves. eu/Members/cclrc/Deliverables/updated-preservation-datastores-interface/at_download/file. pdf.

[16] D24. 1 Report on authenticity and plan for interoperable authenticity evaluation system[EB/OL]. [2015-03-01]. http://www. alliancepermanentaccess. org/wp-content/uploads/downloads/2014/06/APARSEN-REP-D24_1-01-2_5_incURN. pdf.

[17] D24. 2 Implementation and testing of an authenticity protocol on a specific domain[EB/OL]. [2015-03-01]. http://www. alliancepermanentaccess. org/wp-content/uploads/downloads/2014/06/APARSEN-REP-D24_2-01-2_3_incURN. pdf.

[18] CRMdig: A generic digital provenance model for scientific observation[EB/OL]. [2015-03-01]. http://www. cidoc-crm. org/docs/CRMdig-TAPP11. pdf.

[19] SCAPE website[EB/OL]. [2015-03-01]. http://www. scape-project. eu/.

[20] Withers D, Paton N. Design of provenance [EB/OL]. [2015-03-01]. http://www. scape-project. eu/deliverable/d7-1-design-of-provenance-component.

[21] Weise A, Hasan A, Hedges M, et al. Managing provenance in iRODS[EB/OL]. [2015-03-01]. http://link. springer. com/chapter/10. 1007%2F978-3-642-01973-9_75.

[22] Kashi N, Sherwinter N. AV data model: Final specification [EB/OL]. [2015-03-01]. https://prestoprimews. ina. fr/public/deliverables/PP_WP2_D2. 1. 3_AV_Data_Model_R0_v1. 00. pdf.

[23] Mayernik M S, DiLauro T, Duerr R, et al. Data conservancy provenance, context, and lineage services: Key components for data preservation and curation[J]. Data Science Journal, 2013, 12: 158-171.

[24] 李文燕,吴振新.起源信息模型及标准PROV的研究分析[J].情报理论与实践,2015,38(4):23-29.

[25] Assessment of UKDA and TNA compliance with OAIS and METS standards [EB/OL]. [2015-03-01]. http://www.webarchive.org.uk/wayback/archive/20140615012529/http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf.

[26] Provenance management[EB/OL]. [2015-03-01]. http://www.taverna.org.uk/documentation/taverna-2-x/provenance/.

[27] [2015-03-01]. http://www.exlibrisgroup.com/offices.htm.

[28] Missier P, Ludäscher B, Dey S, et al. Golden trail: Retrieving the data history that matters from a comprehensive provenance repository[J]. International Journal of Digital Curation, 2012, 7(1): 139-150.

[29] METS profiles[EB/OL]. [2015-03-01]. http://www.loc.gov/standards/mets/mets-profiles.html.

[30] CCSDS 661.0-R-1, XML formatted data unit (XFDU) structure and construction rules[S]. Washington: CCSDS, 2007.

[31] Dunckley M, Ronen S, Henis E A, et al. Using XFDU for CASPAR information packaging[J]. OCLC Systems & Services: International Digital Library Perspectives, 2010, 26(2): 80-93.

**作者贡献说明:**

**吴振新**:构思论文选题,进行论文修改;

**李文燕**:进行文献调研、资料整理,撰写论文初稿并修改。

## The Application and Research of Data Provenance Technology within Long-term Data Preservation

Wu Zhenxin[1] Li Wenyan[1,2]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190

[2]University of Chinese Academy of Sciences, Beijing 100049

**Abstract**: [**Purpose/significance**] This paper combines the content of provenance and the features of the data preservation, makes a comprehensive study in provenance application within long-term data preservation, and provides a reference for information systems of data preservation to organize and manage the provenance. [**Method/process**] This paper analyzes the explanations of provenance of the relevant standards such as OAIS, PREMIS and TRAC, and makes a comparative study of the application in the existing long-term preservation systems. [**Result/conclusion**] The results is a provenance application framework in data preservation, which summarizes the contents of provenance, and the method to capture, organize, storage and encapsulate provenance.

**Keywords**: provenance event long-term preservation preservation cycle practice

(上接第110页)

## Design and Implementation of Academic Relation and Visualization System

Liu Yuqin[1] Wang Xuefeng[2] Lei Xiaoping[3]

[1]Academic of Printing and Packaging Industrial Technology, Beijing Institute of Graphic Communication, Beijing 102600

[2]School of Management and Economics, Beijing Institute of Technology, Beijing 100081

[3]Institute of Scientific and Technical Information of China, Beijing 100038

**Abstract**: [**Purpose/significance**] Design and implement an academic relation and visualization systerm ItgInsight, to make up for the shortage of the scientific text mining and visualization research. [**Method/process**] ItgInsight has been developed using C# and WPF for constructing and viewing academic relation. Technologies such as data cleaning by field mapping, relations building based on the co-occurrence matrix and association matrix, network diagram and heat map visualization, are used. [**Result/conclusion**] ItgInsight can be used to conduct data cleaning, subject identification, relationship building and visual representation in Chinese or English as far as patent, paper and report. The system with independent intellectual property is stable. It has positive significance to improve intelligence analysis software development in China.

**Keywords**: academic relation visualization system design